Journal of Theoretical and Applied Information Technology
31st March 2014. Vol. 61 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

FRAUD DETECTION IN CARD NOT PRESENT TRANSACTIONS BASED ON BEHAVIORAL PATTERN

[1]RENUGA DEVI T, [2]RABIYATHUL BASARIYA A, [3]KAMALADEVI M
[1, 2, 3] Asst. Professor, School of Computing, SASTRA University, India

E-mail: [1]renugadevi@cse.sastra.edu, [2]basariya@cse.sastra.edu, [3]kamaladevi@cse.sastra.edu

ABSTRACT

Rapid advancements in technology have imperative effect on consumerism. Technological advancements change the way the world operates. The shifting of customer buying patterns is made viable and sustainable through credit cards. Internet merchants are skeptical on Card Not Present transactions. Every advancement has its own intrinsic worth and frailties. Millions of credit card transactions are processed each day. Increases in online shoppers in turn provide more opportunities for credit card usage, which is directly proportional to commit deception. The dearth of tracking fraudulent credit card transactions is due to the increase in white-collar criminals. The way out to this problem is tracked by the behavioral pattern of the customer by implementing the Classifiers Naïve Bayesian and Random Forest to predict the legitimate and fraudulent patterns. The proposed method builds personalize and aggregate model to predict deceitful transactions. Since individual's transactional behavior varies from one another, there comes the need for personalize and aggregate model. The aggregate model performs better than personalized model. Naive Bayesian approach attains best results for personalized model and Random Forest attains best results for aggregate model.

Keywords: *Transactional Behaviour, Fraud Patterns, Naive Bayesian Classifier, Random Forest Classifier, Accuracy, Card Not Present Transactions (CNPT).*

1. INTRODUCTION

Due to the growth of modern technology the mode of payment of an individual has changed significantly. The use of credit cards had become popular and is inseparable in day-to- day activities. At the same time credit card frauds pose a serious threat [10] [11] [12] for issuer as well as the cardholder. In every year the issuer suffers loss in millions. Besides money, the trust between end users and the issuer is weakened. The transactions through credit card can be broadly classified into card present and card not present transactions. The fraudulent issues related to card present transactions are tracked faster as soon as the physical absence of the card. The possibility of fraud committed through CNPT [13] through online and telephone has always shown an upward trend as compared with other types of physical fraud. Both card issuers and cardholders are interested in finding countermeasures to fight fraud. With growing number of transactions, it is very difficult to process the massive amount of data efficiently.

Data mining techniques is the best approach to explore huge data effectively to produce tangible results. Data mining can play a major role in all types of organisations like finance, IT, sales and services etc. There are different mining techniques used to uncover the hidden information from available structured and unstructured data. Each and every individual have different transactional behaviour. Identifying the behavioural vector of a customer's purchase pattern is considered to predict new transactions as dishonest or legitimate. Predictions of new transaction with the aggregate behavioural pattern of customers tend to have promising results compared with prediction based solely on individual behaviour.

Few issues and risk factors in data collection of CNPT are data taken for processing is sensitive and confidential, processing cost beside loss incurred are to be considered. Classification using random forest is effective in aggregate model compared with other classification methods, including SVMs, logistic regression and KNN [15].

In this work the two approaches Naive Bayes classifier and random forest is implemented

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and tested for personalised model and aggregate model with large data set.

1.1 Naive Bayes classifiers

A straightforward probabilistic classifier based on strong naive assumptions. It is a more independent feature model. It proceeds with an assumption that the existence of a particular feature in a set is not based on the existence and nonexistence of other features in the same set. Its inherent probabilistic nature leads to efficient trained supervised learning. In many complex real life situations the naïve Bayes parameters uses the maximum likelihood estimation (MLE).The added advantage of the Bayes is it efficiency in classifying the parameters with less amount of training data due to its assumption of independent variables(conditionally independent),only the variance of each class to be determined. Bayes calculates the prior probabilities based on previous transactions often used to predict the result. Second is to classify the new transaction .The likelihood probability of the new transaction based on parameters is calculated. The prior and likelihood properties are used to calculate posterior probabilities.

The working of Naïve Bayesian Algorithm is explained as follows.

- Naïve Bayesian classification is used for predicting the nature of an unknown class.
- It works based on Bayes Theorem.
- Based on the prior event's knowledge, Bayes theorem will predict future events.

$$P(X|Y) = (P(Y|X) * P(X)) / P(Y)$$

P (X): Hypothesis X's prior probability
P (Y): Training sample Y's prior probability
P (X/Y): Probability of X given Y
P (Y/X): Probability of Y given X

Posterior = (Prior * Likelihood) / (Evidence)
Prior probability for Z = (Number of Z objects) / (Total number of objects)
Likelihood of Y given Z = (Number of Z in the vicinity of Y) / (Total number of Z cases)

With small amount of training data, the parameters can be estimated easily by Naïve Bayesian classifier. It is used to solve both binary classification problem and multiclass classification problem.

1.2 Random Forest Method

Random forest is assembly of decision tress each with a predicted output. Each tree is generated using random set of data. It uses divide and conquer approach to predict the new outcome. Each tree is built on a subset of data to predict the value. The collections of all such weak learners are used to ensemble strong learner. Random forest classifier builds many decision trees by sampling with replacement. In each decision tree, random set of features are selected to determine the splitting criterion and the trees are grown. Each tree will vote for a class. Class, which gets majority votes from each decision tree, will be the final classification output. This classifier is especially suitable when the training set is large.

- The number of classifier variables are M and the number of training cases is N
- Number of input variables considered for making decision is m (m< M).
- A training set is chosen with all N training cases. Choose a sample to be the test set and determine the accuracy of prediction by predicting the classes.
- Randomly choose m decision variables based on which decision is made for each node.
- Build the complete tree without pruning.

The new sample to be predicted is moved down the tree towards the terminal node and the label of the node it ends up is considered as its classifier. This is done repeatedly for all the trees in the ensemble and average classification of all trees is considered as final prediction.

2. LITERATURE SURVEY

Many researchers handle the credit card fraud detection crisis using different mechanisms, supervised model or unsupervised model.Nimisha et al [1] introduced unsupervised technique by dynamically comparing every transaction with user profile and generating the warning for mismatched transactions. FP (Frequent pattern) growth algorithm is used to reveal hidden association rule from the operated transactions to build the Frequent Pattern tree. Zhang yongbin et al [2] proposed an unsupervised model using detection algorithm to compare new transaction against the transaction history. Data pre-processing is applied before detecting the fake transactions. Pre-processing involves collection and transformation of

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

transaction data, uses Fuzzy logic for measuring the transaction activities. As a result of detection algorithm legal and suspected behaviours are separated. Another approach [3] is for identifying the different types of transaction based on past history by combining Bayesian learning and rule based filtering.

V. Dheepa et al [4] projected the method for detecting the credit card fraud using support vector machine, it is an emerging trend for solving the classification related problem [5]. Another method proposed by Venkata ratnam et al [6] for detection is data stream outlier detection algorithm using k-nearest neighbours.This outlier indicates the unusual transaction characteristics and the transaction incidence. Neural network [7] is used for classifying the transaction behaviour as different clusters from low to very high risk, uses SMNNN technique for classifying into different group of clusters. Shailesh [8] suggest the Hidden Markov model for handling the same, HMM is standard tool for solving various problems. Many researchers thrash out the revise of similar available card detection mechanisms [9]. Rong-Chang Chen et al [14] proposed a personalized method for fraud identification using SVM and Neural Networks specifically BPN (Back Propagation Network) in order to handle the different behaviours of credit card holder. Pre-processing of data received in transaction level in the form of data aggregation allows to handle heterogeneous multidimensional data effectively by taking into account the necessary fields to develop aggregate model for identifying suspicious transactions [15] .The way business and financial institutions take more effort to prevent fraud is effective when the two main issues time efficiency and cost savings are considered [11]. Different types of frauds are as many as various financial institution products and services. [16]. In this work the performance of Bayes classifier and random forest is tested on aggregate and personalised model.

3. PROPOSED METHOD

Proposed method uses behavioural data mining technique to detect fraud. Since transaction behaviour varies from one individual to another, fraud detection is analysed here by building two different models Personalised and aggregate. Personalised model deals with tracking every individual's current transaction based on the past transactional behaviour of the respective individual. It identifies the purchase patterns of customer based on the transaction set variables assumed below. Each new transaction is analysed based on the set of assumed variables on the individual's history using random forest and naive Bayes algorithm.

Aggregate model deals with tracking every individual's current transaction based on the past transactional behaviour of all individuals. It decides the purchase pattern of the customer by considering the below given characteristics of the transaction with all other customers using random forest and naive Bayes.

The transactional set is collected in the form of online questioner. The data collection includes:

Credit card number: the sixteen digit unique number of each cardholder.

Transaction number: the unique number generated for each transaction.

Expiry date of the card: the date up to which the credit card is valid.

Merchant category: the various merchant categories in which the cardholder may purchase goods.

Volume of purchase: the quantity of goods purchased by the cardholder.

Currency Information: specifies whether the cardholder has informed about the changes in currency in situations like travelling overseas.

Amount of transaction: amount spent by the cardholder for the purchase made by him/her.

Time of transaction: time at which the cardholder makes transaction (system time considered).

Frequency of transaction: the number of transactions made in a specified period of time.

Credit Score: suspicious score calculated for each individual's transaction

The information collected will act as the training set for both models. Considering these factors as the basic metric, the current transaction is compared with the history of particular individual in case of personalized model and with the history of all individuals in case of aggregate model thereby suspecting the nature of the transaction.

Journal of Theoretical and Applied Information Technology
31st March 2014. Vol. 61 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

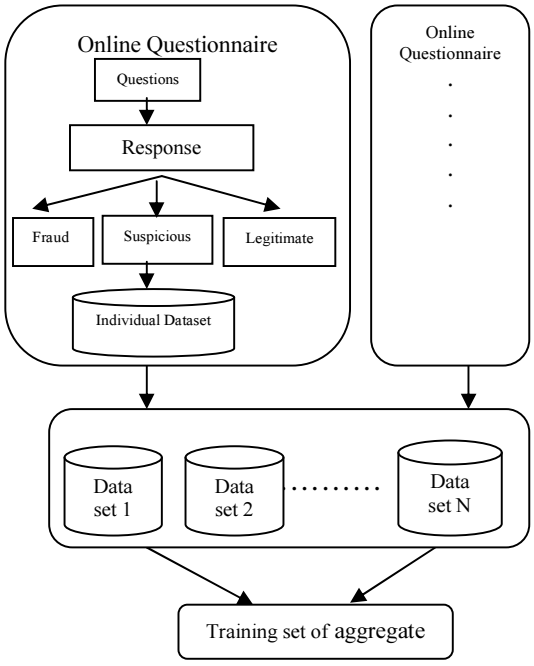

Figure 1. Personalised Model And Aggregate Model

Advantages of proposed methods are as follows. This study focuses on the performance of algorithm on each model. Thus by comparing the performance of both the models, it is found which model outperforms the other in detecting fraud more accurately thereby eliminating the need for constructing the other. Moreover the factors that are used here like credit score helps to detect fraud more accurately.

The training set and the test set used for building models includes legitimate transaction, fraud transaction as well as suspicious transactions. The diagrammatic representation of how personalized and aggregate models are built is shown in Figure 1

Sample Test set

The current transaction details entered in the test set are as follows:

- Credit card number: 1234567890123456
- Transaction number: 678954
- Expiry date of the card: 20.12.2020
- Merchant category: Clothing and lifestyle
- Volume of purchase: 5
- Currency Information: Yes
- Amount of transaction: 9000
- Time of transaction: 10:00:00
- Frequency of transaction: 5

Each of the above fields is validated. The data collected is pre processed before processing. After these transaction details are filled, credit score for this transaction is calculated. According to the importance of the factors used for detecting fraud, weightage will be assigned to each of these fields. Credit score is calculated by considering the parameters collected in the questioner. The flag values for each attribute calculated based on the importance of the attribute using information gain.

$$Score = (15*flagexp) + (1.5*flagmer) + (2*flagvol) + (2*flagdis) + (20*flagcur) + (3*flagamn) + (3*flagtim) + (2.5*flagfrq)$$

Score ranges from 0 to 50.This set acts as the test set to the classifier for which the class is predicted. Based on credit score, legitimate and fraud transactions are identified

4. EXPERIMENTAL RESULTS AND ANALYSIS

The ultimate aim of this study is to analyse the performance of personalised and aggregate model and the efficiency of random forest and Naive Bayesian classification on the same. Two personalized model using Naive Bayesian classifier as shown in fig 2 and fig 3 and one aggregate model as shown in fig 4.Two personalised model using random forest as shown in fig5 and fig 6 and one aggregate model as shown in fig7. These experiments are carried out using WEKA data mining tool.

The following factors play a major role in analysing the performance. They are listed as follows:

-Accuracy
-Precision
-Recall
-F Measure
-True Positive
-False Positive.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table1. Performance of Personalized Model

Model	Personalized model -I		Personalized Model II	
Metrics	NB	RF	NB	RF
Accuracy (%)	73.42	57.69	73.80	54.85
Precision	0.7118	0.5504	0.6956	0.5092
Recall	0.4844	0.4759	0.4683	0.4624
F-Measure	0.5765	0.5104	0.5597	0.4846
TP	0.7115	0.5504	0.6956	0.5092
FP	0.288	0.4495	0.3043	0.4907

Analyzing the experiment based on above factors clearly show that the Naïve Bayesian is best for Personalized model in detecting fraud accurately when compared to Random Forest classifier as given in table I.

Table2. Performance of Aggregate Model

Metrics	NB	RF
Accuracy (%)	66.51	82.33
Precision	.625	.7777
Recall	.4725	.4714
F-Measure	.5381	.5869
TP	.625	.7777
FP	.375	.2222

For aggregate model, it is understood clearly that Random forest classifier detects fraud more accurately than Naïve Bayesian classifier since the data set is large in aggregate model as in table II.

Table3. Personalized Vs. Aggregate Model

Model	NB	RF
Personalized I	73.42	57.69
Personalized II	73.80	54.85
Aggregated	66.51	82.33

From table III the performance of aggregate model is better than personalized mode. This eliminates the need for building personalized model, which saves extra cost. Building personalized model for each and every cardholder is practically impossible since we have to maintain separate database for each and every individual.

5. CONCLUSION

The physical experience of in store purchase is irreplaceable for excellent producer consumer relationship and understanding the need and behaviour of end user. E-commerce always has ripple effect for changing customer's behaviour. The comparative study on the creation of personalized and aggregate model reveals that aggregate model outperforms personalized model in detecting fraudulent transactions. Random Forest best suits than the Naïve Bayes for the aggregate model and Naïve Bayesian best suits personalized model. Since the size of the training dataset is larger for aggregate model when compared to personalized model, the former result was achieved. The credit score of a transaction is one of the most important factor used to predict the transaction. Moving a step ahead, measures should be taken to prevent fraud once it is detected. As an extension of this work alert message can be sent for suspicious transactions. The work can also be extended for card present transactions by learning customers purchase pattern along with additional parameters like customers bank transactional information, location of card usage with respect to its location of issue.

REFERENCES

[1] Nimisha Philip, Sherly K.K, "Credit Card Fraud Detection Based on Behaviour Mining" *TIST.Int.J.Sci.Tech.Res.,* Vol.1 , 2012, pp. 7-12.

[2] Yongbin Zhang, Fucheng You, Huaqun Liu, "Behavior-Based Credit Card Fraud Detecting Model" *Fifth International Joint Conference on INC, IMS and IDC*, 2009.NCM, pp. 855-858.

[3] Suvasini Panigrahi, Amlan Kundu, Shamikr Sural and A.K.Majumadar,"Credit Card Fraud Detection: A Fusion Approach Using Dempster Shafer Theory and Bayesian Learning" *Information Fusion,* 10, 2009,pp. 354-363.

[4] Dheepa.V, R. Dhanapal, "Behavior Based Credit Card Fraud Detection Using Support Vector" *ICTACT Journal on Soft computing*, 02, 2012, pp. 391-397.

[5] Zan Huang, Hsinchun Chen, Chia-Jung Hsu , Wun-Hwa Chen , Soushan Wu, "Credit Rating Analysis with Support Vector Machines and Neural Networks: A Market Comparative Study" *Elsevier-Decision Support Systems,* Vol. 37,2004, pp. 543-558.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[6] Venkata Ratnam Gnnji, Siva Naga Prasad Mannem, "Credit Card Fraud Detection Using Anti-k Nearest Neighbour Algorithm" *International Journal on Computer Science and Engineering (IJCSE)*, 4, 2012,pp. 1035-1039.

[7] Shailesh S. Dhok, "Credit Card Fraud Detection Using Hidden Markov Model", *International Journal of Soft Computing and Engineering* (IJSCE), 2,2012, pp. 88-92.

[8] Benson Edwin Raj S, A, Annie Portia, "Analysis on Credit Card Fraud Detection Methods" IEEE *International Conference on Computer, Communication and Electrical Technology* (ICCCET2011), 18th & 19th March, 2011,152-156.

[9] Barry Masuda "Credit Card Fraud Prevention: A Successful Retail Strategy" *crime prevention*, 1986, pp. 121-134

[10] Linda Delamaire, Hussein Abdou, John Pointon, "Credit Card Fraud and Detection Techniques: a Review" *Banks and Bank Systems,* 4, 2009,pp. 57-68.

[11] Tej Paul Bhatla, Vikram Prabhu, Amit Dua, "Understanding Credit Card Frauds" *Cards Business Review*, Tata Consultancy Services, 2003.

[12] Richard J. Bolton and David J. Hand, "Statistical Fraud Detection: A Review" *Statistical Science,* 17, 2002, pp. 235–255.

[13] Rong-Chang Chen, Shu-Ting Luol, Xun Liang, Xun Liang, "Personalized Approach Based on SVM and ANN for Detecting Credit Card Fraud" *International Conference on Neural Networks and Brain,* 2,2005, pp. 810 – 815.

[14] Whitrow C, D J Hand, P Juszczak, D Weston, N. M. Adams, "Transaction Aggregation as a Strategy for Credit Card Fraud Detection" *Data Mining and Knowledge Discovery*, 2009,30-55.

[15] Anderson, R, "The Credit Scoring Toolkit: theory and practice for retail credit risk management and decision automation" *New York: Oxford University Press*, 2007.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
Naive Bayesian Classification
 1234567890123456       660380  true    Healthcare      Yes     low     far     normal  night time      less
pgivenl is 0.0923521
pgivenf is 7.575758E-4
pgivens is 0.0
Credit score for current transaction is 5.0
pll is 0.030784033
plf is 0.0
pls is 0.0
ptotal is 0.030784033
plgiven is 1.0
pfgiven is 0.0
psgiven is 0.0

Class Of Entered Tuple is Legitimate
```

Figure 2. Predicting current Transaction of Personalized model 1 using Naive Bayesian Algorithm

```
Naive Bayesian Classification
 9876543210123456       691812  false   Games   Yes     high    far     normal  night time      more
pgivenl is 0.0
pgivenf is 2.1768712E-4
pgivens is 0.0
Credit score for current transaction is 24.5
pll is 0.0
plf is 1.8140594E-5
pls is 0.0
plgiven is 0.0
pfgiven is 1.0
psgiven is 0.0

Class Of Entered Tuple is Fraud
```

Figure 3.Predicting current Transaction of Personalised model 2 using Naïve Bayesian Algorithm

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
Naive Bayesian Classification
1234567890123456        324612  true    Oil and Gas     No      low     near    normal  night time      less
pgivenl is 0.0
details 1.0     1.0     0.88235295      0.7368421       0.0     1.0     0.1724138       0.96428573
details 0.15789473      0.3     0.8     0.6363636       0.0     0.3846154       0.38709676      0.26190478
pgivenf is 0.0
details 1.0     0.35    1.0     0.4375  1.0     1.0     1.0     0.45238096
pgivens is 0.069270834
Credit score for current transaction is 23.0
pll is 0.0
plf is 0.0
pls is 0.015874567
ptotal is 0.015874567
plgiven is 0.0
pfgiven is 0.0
psgiven is 1.0

Class Of Entered Tuple is Suspicious
```

Figure 4. Predicting current Transaction of Aggregated model Using Naïve Bayesian Algorithm

```
1234567890123456,324612,true,'Oil and Gas',low,near,No,normal,'night time',less,23,'?'

Random forest of 10 trees, each constructed while considering 3 random features.
Out of bag error: 0.4167


 Predicting TestSet

1.0 -> Fraud
```

Figure 5. Predicting current Transaction of Personalized model 1 Using Random Forest Algorithm

```
9876543210123456,249824,true,Education,low,near,Yes,normal,daytime,less,0,'?'

Random forest of 10 trees, each constructed while considering 3 random features.
Out of bag error: 0.3333


 Predicting TestSet

0.0 -> Legitimate
```

Figure6.Predicting current Transaction of Personalised model 2 Using Random Forest Algorithm

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
1234567890123456,324612,true,'Oil and Gas',low,near,No,normal,'night time',less,23,'?'

Random forest of 10 trees, each constructed while considering 3 random features.
Out of bag error: 0.2292


 Predicting TestSet

2.0 -> Suspicious
```

Figure7. Predicting current Transaction of Aggregated model using Random Forest Algorithm